Geophysical Journal International

Geophys. J. Int. (2023) **234,** 2502–2519
Advance Access publication 2023 June 24
20GJI Seismology

https://doi.org/10.1093/gji/ggad256

Operational Earthquake Forecasting in Italy: validation after 10 yr of operativity

I. Spassiani,[1] G. Falcone,[1] M. Murru[1] and W. Marzocchi[2]

[1]*Istituto Nazionale di Geofisica e Vulcanologia (INGV), Via di Vigna Murata,* 605, 00143 *Rome, Italy. E-mail: ilaria.spassiani@ingv.it*
[2]*Department of Earth, Environmental, and Resources Sciences, The University of Naples Federico II, Via Cinthia,* 21–80126, *Naples, Italy*

Accepted 2023 June 21. Received 2023 May 19; in original form 2023 February 7

SUMMARY

In this paper, we gather and take stock of the results produced by the Operational Earthquake Forecasting (OEF) system in Italy, during its first 10 yr of operativity. The system is run in real-time: every midnight and after each M_L 3.5+ event, it produces the weekly forecast of earthquakes expected by an ensemble model in each cell of a spatial grid covering the entire Italian territory. To evaluate the performance skill of the OEF-Italy forecasts, we consider here standard tests of the Collaboratory for the Study of Earthquake Predictability, which have been opportunely adapted to the case of the overlapped weekly OEF forecasts; then we also adopt new performance measures borrowed from other research fields, like meteorology, specific to validate alarm-based systems by a binary criterion (forecast: yes/no; occurrence: yes/no). Our final aim is to: (i) investigate possible weaknesses and room for improvements in the OEF-Italy stochastic modelling, (ii) provide performance measures that could be helpful for stakeholders who act through a boolean logic (making an action or not) and (iii) highlight possible features in the Italian tectonic seismic activity.

Key words: Probabilistic forecasting; Statistical methods; Time-series analysis; Earthquake interaction, forecasting, and prediction; Statistical seismology.

1 INTRODUCTION

In 2009, the International Commission for Earthquake Forecasting (ICEF) for Civil Protection, nominated by the Italian government after the M_L 5.9 (M_w 6.1) earthquake occurred in L'Aquila (Italy) on 2009 April 6, introduced the concept of Operational Earthquake Forecasting (OEF) as a tool to gather and disseminate authoritative information about the time dependence of seismic hazards, as well as to help both stakeholders to establish rational seismic risk reduction strategies, and communities to be prepared for potentially destructive earthquakes; the full report was released in 2011 (Jordan *et al.* 2011).

Earthquake forecasting was not a novelty, because models already existed and were applied in practice (e.g. Reasenberg & Jones 1989; Rhoades & Evison 2004; Gerstenberger *et al.* 2005; Holliday *et al.* 2005; Michael *et al.* 2020). Nonetheless, the ICEF report emphasizes the importance of features that were not yet considered at that time, such as, for example, the need to provide continuous information and the importance of testing the forecasts. The first model that fully complies, at least in principle, with the ICEF requirements has been published in 2014 by the seismic hazard centre at the Istituto Nazionale di Geofisica e Vulcanologia (INGV, Marzocchi *et al.* 2014).

Although the model has been published in 2014, since 2009 the OEF-Italy system is run 24/7 on a computer physically placed at the INGV in Rome (Marzocchi & Lombardi 2009; Marzocchi *et al.* 2012, 2017). It is performed over a specific $0.1^\circ \times 0.1^\circ$ grid lattice of $N_c = 8993$ cells purposely placed inside an area covering the whole national territory, opportunely selected for Italy according to the standards proposed by the Collaboratory for the Study of Earthquake Predictability (CSEP, https://scec.usc.edu/scecpedia/CSEP_Working_Group, Schorlemmer *et al.* 2018), which is an international infrastructure aimed at standardizing and quantitatively evaluating earthquake predictions and forecasts at a global scale. Both the CSEP polygon and the OEF grid are shown in Fig. 1, where it can be observed that Sardinia is not included in the analysis as instrumental seismicity in this island is not enough to calibrate the model: this area is in fact not affected by active tectonics like in the case of the Apennines or other Italian regions. Although not visible in the map, we stress that the area around the Etna volcano (Sicily) is also excluded from the analysis, because the seismic activity beneath active volcanic areas is driven by different mechanisms (e.g. magma intrusions), which are not well captured by the probabilistic models behind OEF-Italy.

In the specific, at the midnight of each day, and after the occurrence of any M_L 3.5+ event recorded in real-time by the Italian Seismic Network, the OEF system produces the next week's forecast of earthquakes with local magnitude M_L 4.0+ or 5.5+, or with microseismic Modified Mercalli Intensity *MMI* VI+, VII+ or VIII+, fixed at user's will. In what follows, we will refer to these

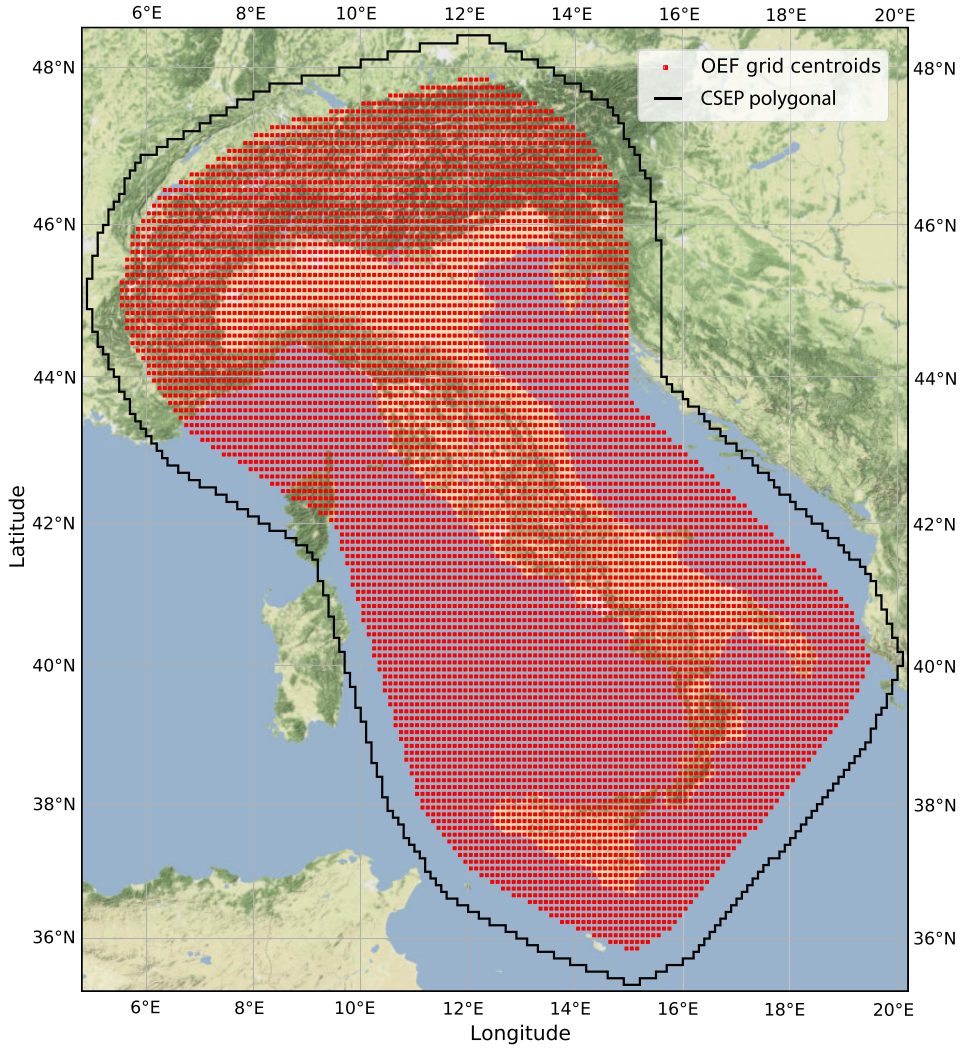

Figure 1. CSEP polygon (piecewise black curve) embracing the Italian territory, and centroids (red points) of the $0.1^\circ \times 0.1^\circ$ grid lattice over which the OEF experiment is performed.

selected earthquakes as 'target events'. The choice of weekly time windows has been agreed upon with the Italian Civil Protection for practical utility, but we stress that the system is able to provide shorter-term forecasts (e.g. daily). The flexibility of the OEF-Italy system to deliver forecasts relative to different timescales is indeed an important point to underline. In general, the proper length of the forecasting window is related to the use those forecasts are intended to. A short timescale could be appropriate for (quasi-)daily information, or in case of emergency, when a precise decision-making table is needed (e.g. during an ongoing seismic sequence). Longer time periods are instead crucial for the building code, for example in the reconstruction process of damaged municipalities. In-between forecasting windows may meet different requirements, mostly related to cases of no seismic emergency, but still when the forecast is required for a near-future (e.g. when the state of emergency is closed, but the restoration process is still ongoing). These stages may also overlap, and their duration depends on several governmental and socioeconomic factors, as well as on the full extent of the disaster (Michael 2012). That said, the 1-week time window is been used these past years by the Italian Civil Protection for reasons of internal organization, therefore these are the forecasts we test in this paper. The specific length of 7 d has implied a 'practical use' of the delivered forecasts mostly during an emergency, or to monitor the evolution of an ongoing seismic sequence; still, the discussion among the parties involved is continuous, such to ensure that the system is prepared based on any different need.

The OEF-Italy forecasts are computed from mathematical models well-known in statistical seismology (see Section 1 in the Supp¬rting Information), and are released through time-dependent seismic maps coloured according to the obtained expected probabilities (e.g. see the map in Fig. 2). In addition, the system produces and stores the expected rates calculated from each model at each run.

The graphical interface of OEF-Italy consists of an embedded Leaflet Map reflecting the current weekly probability of the target events inside a spatial window (circle or rectangle), directly selected by the user from the interactive dashboard. An additional interactive graph shows the temporal evolution of the weekly probabilities, which can be zoomed on any time interval of interest. The probability value of the last run produced by the system is also shown. An example of the OEF interface is given in Fig. 2, where we selected M_L 4.0+ target events within a small rectangle in Central Italy (black box in the map).

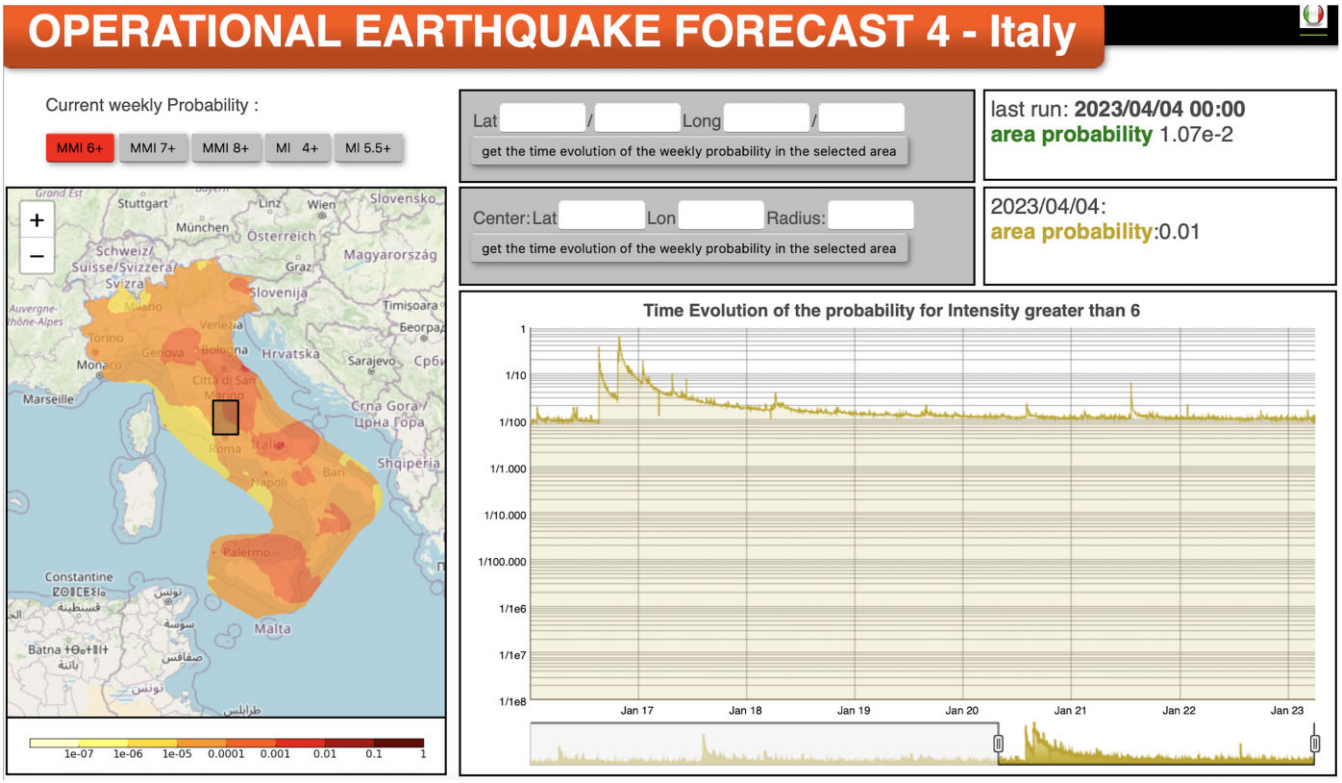

Figure 2. Graphical interface of the OEF-Italy system, example for a small rectangular area in Central Italy (black box in the map). On the left: embedded Leaflet Map of the current weekly probability for the selected area, where we selected the events M_L 4.0+ as target. On the right: timeline of the probability history from 2009 to 2021 January (bottom), together with two boxes (top right) showing the probability of the last run, and the probability computed at the date pointed by the cursor along the timeline.

To date, the OEF-Italy system is not open to the public and the time-dependent seismic information has been released in structured manner since 2015 only to the Major Risk Commission of the Italian Civil Protection in four-months reports and in daily reports during important seismic sequences, or upon specific request. Still, the system is running in real-time, and the possibility of continuously accessing the flow of information produced is being discussed, as well as the possibility of spreading results to the public, together with the best way to disseminate them (e.g. see Michael *et al.* 2020; Becker *et al.* 2020, for reference). Major obstacles come out from a legal system which is unclear on roles and responsibilities of scientists involved in delivering this information. This issue has still easy-to-predict major consequences in Italy after the infamous L'Aquila earthquake trial (Marzocchi 2012).

In this work, we gather and take stock of the results obtained for the first 10 yr of the OEF experiment in Italy, to the aim of assessing the reliability of the forecasts produced, in comparison with the real earthquake catalogue recorded in the same period. The analysis in this paper is performed by means of several statistical tests, so as to give a clue on possible improvements in the OEF-Italy system and insights on Italian seismic activity. Since the single earthquake forecasting models are continuously under evaluation in CSEP experiments (e.g. Taroni *et al.* 2018), here we provide an additional contribution analysing the time overlapping ensemble forecasts through different and complementary statistical approaches, that are widely used in different fields.

We finally stress that only the OEF-Italy rate and probability forecasts are tested here, but a similar study can be done to test the Modified Mercally Intensities delivered by the system. In fact, as explained before, the system contains also the modelling of seismic-wave attenuation with distance from the source and of the site effects (Marzocchi *et al.* 2014). This will be object of a future study, more focused on testing the OEF-Italy performance in respect to the building code and the risk scenarios.

2 THE FORECASTING SET UP AND THE EARTHQUAKE CATALOGUE

Although the earthquake forecasting experiments started in 2009, the OEF system became real-time operative in Italy on 2013 January 1. This study is therefore carried out in the time window 2010–2020, but only from 2013 January 01 until 2020 May 26 the tests are purely prospective; before 2013, the tests are carried out in a pseudo-prospective mode (all parameters have been set up using the data before 2010). The probabilistic forecasts are obtained from an ensemble model, which is the weighted combination of three versions of three stochastic models typically used in statistical seismology: the Epidemic Type Aftershock Sequence (ETAS) proposed by Lombardi & Marzocchi (2010), the Epidemic Type Earthquake Sequence (ETES) by Falcone *et al.* (2010) and the Short-Term Earthquake Probability (STEP) by Woessner *et al.* (2010). For further details, see Section 1 in the Supporting Information.

As specified above, the weekly forecasts are provided by the OEF-Italy model every day at midnight and after each earthquake of magnitude 3.5 or larger, for each cell of the spatial grid covering the Italian territory with cells of $0.1^\circ \times 0.1^\circ$ (see Fig. 1).

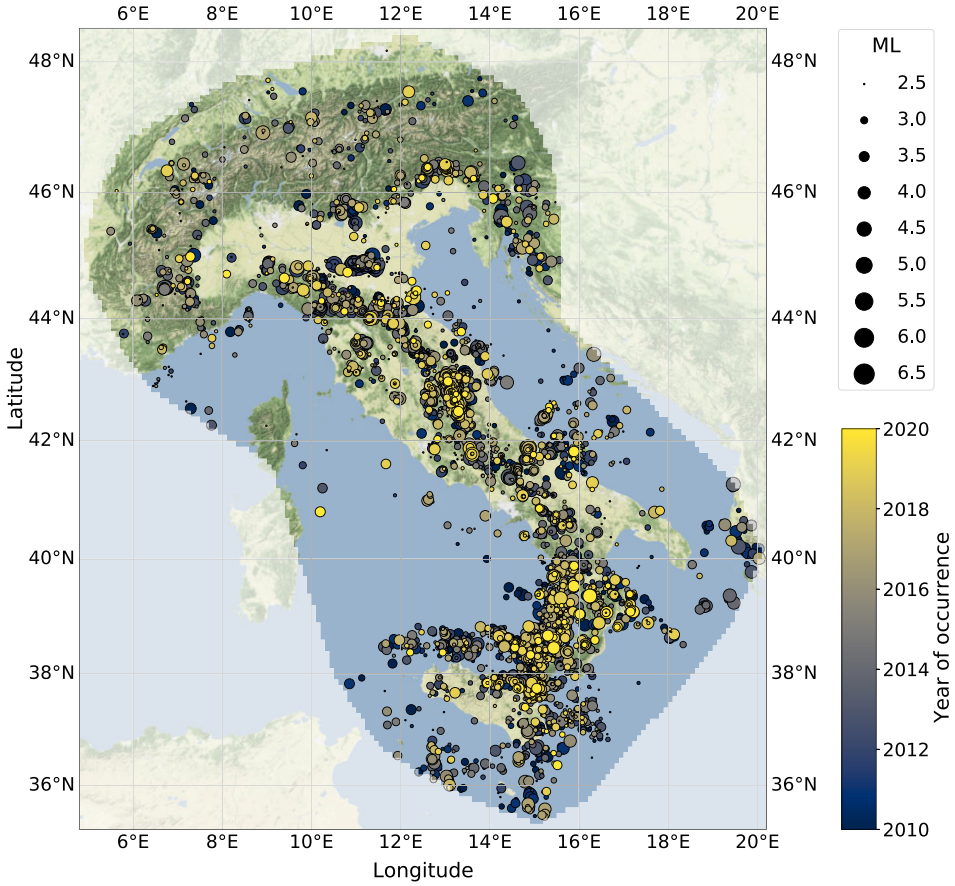

Figure 3. Map of the seismic events recorded in Italy during the years 2010–2020. The size increases with the events' magnitude, while the colour varies with the temporal occurrence (blue to yellow for less to more recent events). The map is masked by the CSEP polygon (see Fig. 1).

The target earthquakes we consider here are the shallow events (depth $\leq$30 km) with $M_L \geq 4.0$ that occurred in the grid within the testing temporal interval previously defined. The number of total OEF-Italy time forecasts is $N_t = 3407$.

The earthquake catalogue recorded by the INGV Italian Seismic Network during the forecasting window 2010–2020 inside the CSEP polygon (Fig. 1) consists of 11 272 events with $M_L \geq 2.5$. They are mapped in Fig. 3, where the dots' size increases with the events' magnitude, and their colour changes from blue to yellow as the occurrence time becomes more recent. A Lilliefors (1969) test identified the value of 2.7 for the completeness magnitude. We stress that this value is hands-down underestimated for different reasons just after the occurrence of a strong earthquake, thus introducing an incompleteness that protracts for a variable time window, usually named short-term aftershocks incompleteness (STAI, Kagan 2004; Lippiello *et al.* 2019).

In the pure prospective experiment starting from 2013 January 01, the number of target events (depth $\leq$30 km, $M_L \geq 4.0$) is 182. Ten of these events occurred at the border of the CSEP polygon, that is in the sea or abroad, indeed outside the grid cells on which the OEF-Italy forecasts are computed. Then, they are not considered in the analysis, leaving a total of 172 target events that are mapped in Fig. 4. They occurred in only 107 d of the tested time-window considered, and inside 87 cells of the spatial grid, 19 of which having more than one target event.

Fig. 4 shows that Central Italy experienced the largest number of target earthquakes, and in fact this area has been interested by a strong earthquake sequence started on 2016 August and characterized by several $M_L \geq 5.0$ events occurred till 2017 January. This is the Amatrice-Norcia-Visso sequence, that began with the M_L 6.0 (M_w 6.0) event occurred on 2016 August 24, with epicentre between the municipalities of Accumuli - Amatrice (Rieti Province, Lazio Region) and Arquata del Tronto (Ascoli-Piceno Province, Marche Region). Two strong shocks followed on 2016 October 26, with M_L 5.4 and 5.9 (M_w 5.4 and 5.9) respectively, near the municipalities of Visso, Ussita and Castelsantangelo sul Nera (Macerata Province, Marche Region). The strongest event of the Central Italy sequence occurred on the next October 30, with M_L 6.1 (M_w 6.5) and epicentre between Norcia and Preci (Perugia Province, Umbria Region). About two months later, and precisely on 2017 January 18, four additional shocks occurred with a moment magnitude higher than 5. This is actually the strongest sequence registered in the spatio-temporal tested window we selected for the analysis.

3 TESTING THE RELIABILITY OF OEF-ITALY FORECASTS

Here, we implement a testing phase that complement the tests carried out in the CSEP framework. In particular, we test heavily overlapped forecasts of strongly clustered events. To minimize the effect of strong clustering, we translate the observations in dichotomous observations, that is, we consider the occurrence or not of at least one target event (depth $\leq$30 km, $M_L \geq 4.0$) in each spatio-temporal

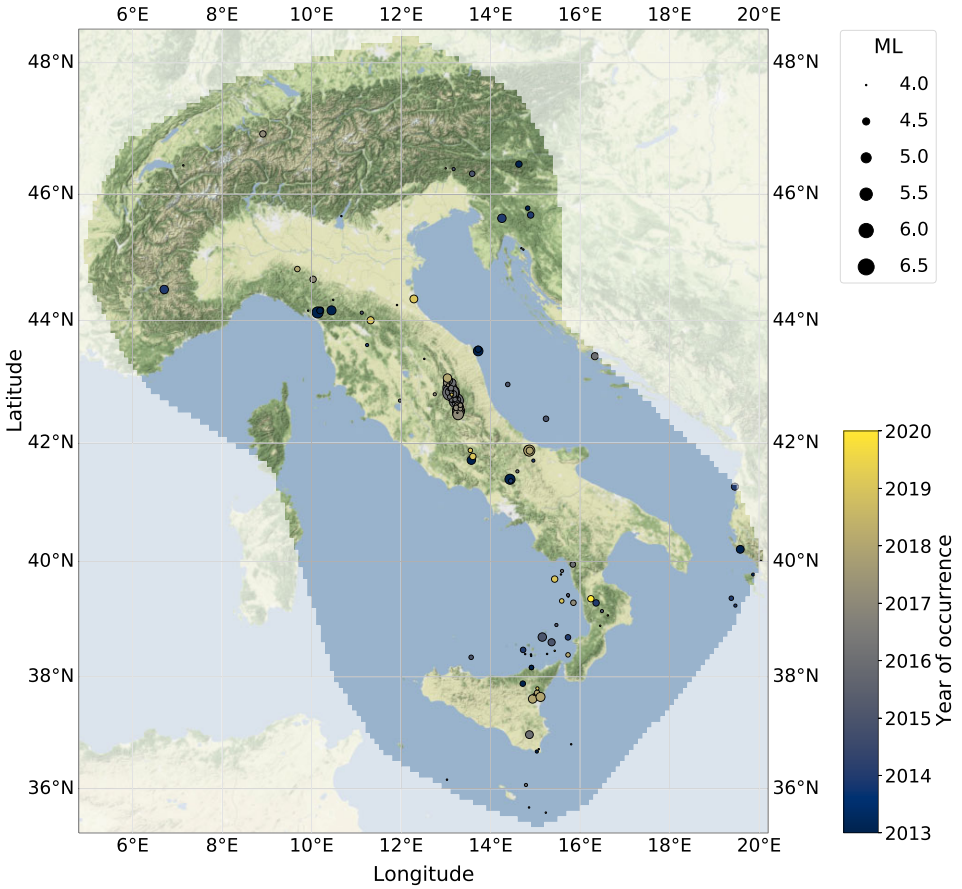

Figure 4. Map of the 172 target events (depth ≤30 km, $M_L \geq 4.0$) recorded in the Italian catalogue in the period 2013 January 01–2020 May 26. The size increases with the events' magnitude, while the colour varies with the temporal occurrence (blue to yellow for less to more recent events). The map is masked by the CSEP polygon (see Fig. 1).

bin (Bayona *et al.* 2022):

$$O_{ij} := \begin{cases} 1, & \text{if } 1+ \text{ target events in cell} C_j \text{ during week from run} i \\ 0, & \text{if } 0 \text{ target events in cell} C_j \text{ during week from run} i. \end{cases} \quad (1)$$

To handle the marked time overlap of the forecasts, we gather all the spatio-temporal bins together in a single, high-dimensional array with forecasts×cells components. This means to flatten the multidimensions of space and time in a unique 'spurious' dimension capturing the general features of the forecasts. More precisely, at time i we associate to the cell j only the ensemble forecast λ_{ij} (rate), or P_{ij} (probability); in other words, for every spatio-temporal bin we analyse each individual rate or probability, without considering that it refers to overlapping weekly time windows. Keeping each individual forecast unavoidably inflates the number of target earthquakes (each individual target earthquake belongs to several time windows). Nonetheless, this approach allows us to optimize the computational costs, to evaluate the OEF-Italy forecasts in their whole, and to better reflect the true course of events when the system is run in real-time.

The reliability of the OEF-Italy forecasts in about the first 7 yr of real-time operativity (precisely, from 2013 January 01 till 2020 May 26) is then assessed through several statistical procedures, briefly summarized in Table S1 of the Supporting Information. As mentioned above, these tests complement the analyses made in CSEP and highlight some interesting and novel aspects of OEF-Italy forecasting reliability.

4 A REVISED N-TEST FOR OVERLAPPING TIME WINDOWS

In this section, we test if the number of target earthquakes is compatible with the forecasts. We produce 10 000 synthetic binary matrices of elements

$$\hat{O}_{ij} := \begin{cases} 1, & \text{if} \quad r_{ij} < P_{ij} \\ 0, & \text{if} \quad r_{ij} \geq P_{ij}, \end{cases}$$

where $(i, j) \in N_t \times N_c$ and $r_{ij} \sim$ Uniform(0,1). Each binary matrix is then compared to the binary matrix $\mathbf{O} = \{O_{i,j}\}$ of observed target events, whose elements are defined in eq. (1). This approach assumes independency among bins, which are actually heavily overlapped. To overcome this issue, we rescale the simulations in the whole temporal window covered by all the 3407 runs by a factor $F = \frac{D_{T,\,\mathrm{all}}}{D_T}$, where $D_{T,\mathrm{all}}$ is the number of days in the total number of individual forecasts, about 7×3407 (this is a lower bound due to the intensification of the forecasts during a crisis), and D_T is the temporal length (in days) between the first and the last runs dates. This factorization allows us to consider the weekly forecasts as if they were actually non-overlapped.

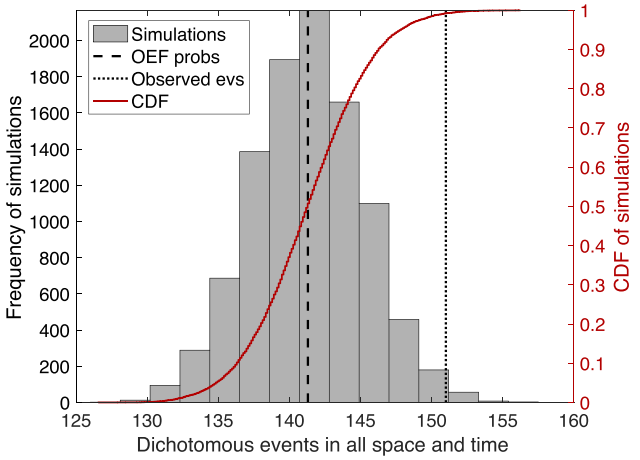

Figure 5. Histogram (grey bars) and CDF (red step line) of the number of spatio-temporal OEF bins with at least one target event, among 10 000 synthetic catalogues. The dashed black line represents the sum of the overall ensemble probabilities, while the dotted black one represents the number of spatio-temporal bins containing at least one observed target event.

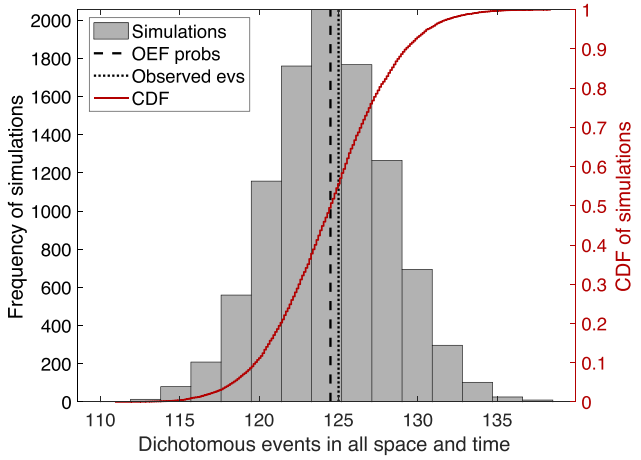

Figure 6. The same as Fig. 5, but removing from the analysis 1 d after each M_L 5.4+ event (6 events in total).

In our case, we find $F = 8.826$. We divide each component of the simulation array $\mathbf{S} = (Sim^{(1)},...., Sim^{(10000)})$ by this factor, where $Sim^{(i)}$ corresponds to the total number of *ones* obtained over all the runs in the *i*th simulation. Finally, we compute the histogram and the cumulative number of **S**, and we compare them to the number of ones in the binary matrix **O** of observed target events. In the whole 3407×8993 bins, we count 151 *ones*. We stress that multiple events within a week are discarded, and that is the reason why we get 151 instead of 172 target events.

The procedure just described is actually a way of adapting to the case of overlapped forecasts the classical *N*-statistical test (Zechar 2010; Zechar *et al.* 2010; Taroni *et al.* 2018). Indeed, we can evaluate the consistency between the number of forecasted earthquakes in all space–time–magnitude bins, and the number of observed target events in a testing spatio-temporal window of interest.

Results are shown in Fig. 5, where we plot the histogram (grey bars) and the cumulative distribution function (CDF, red step line) of the synthetic observations obtained for all the spatio-temporal OEF bins, after the rescaling procedure described above. The results highlight the underestimation of the number of bins with at least one observed target earthquake (black dotted line), which falls in the right tail of the histogram distribution. This is what expected, as the high incompleteness entailed after the strongest shocks of the Central Italy sequence naturally induces an underestimation of the forecasts. Helmstetter *et al.* (2006) have shown that STAI should last, in general, less than 1 d. However, to be conservative, we removed here 1 d after each M_L 5.4+ event (6 events in total) from the analysis, thus strongly reducing this effect, as shown in Fig. 6. We also repeated the analysis by excluding only 6 and 12 hr, and we obtained very similar results.

To further investigate this point, in Fig. 7 we separately deal with time (left-hand column) and space (right-hand column). Top line panels confirm the consistency between simulations and OEF data, both in time and space, while the bottom panels highlight that the real seismicity differs from synthetic data mainly in correspondence of the Central Italy sequence. The forecasts for which the difference is statistically significative, that is, number of observed events that is outside the 99 per cent confidence interval (CI) obtained from simulations, are represented in the bottom left panel by circles coloured in grey scale from black to white as the difference increases. Although this is not clearly visible from the figure, due to the superimposition of the circles corresponding to outliers very close in time, they are 63 (we discard 7 additional outliers ascribed to Etna volcano, see Introduction), and are listed in Table S2 of the Supporting Information. The cells they are associated to are given as circles in the bottom right panel of Fig. 7, and are also mapped in Fig. 8. We further note that, when removing 1 d after each M_L 5.4+ event, the outliers reduce to 41.

We finally focus on two cells involved in the 2016–2017 Central Italy sequence, that is, the ones containing the municipalities of Amatrice and Norcia, respectively in top and bottom line panels of Fig. 9. By looking at the left-hand plots, we note that the bins with at least one observed target event are underestimated only in the case of Norcia (black dotted line outside the histogram bars). Indeed, Amatrice was not preceeded by a strong earthquake activity like in Norcia, where the incompleteness induced by previous seismicity caused the visible underestimation. When we remove from the analysis 1 d after the two events with M_L 5.4+ occurring in the cell with the municipality of Norcia, simulated and observed seismicity become again consistent (see Fig. 10). For both the two cells we also compute the 99th percentile of the 'frequency of at least one target event among the 10 000 simulations'. We find that, 99 per cent of times, this frequency is below the value of 15 per cent for the cell with Amatrice, increasing to about 35 per cent for the cell with Norcia. This is shown in the right-hand panels of Fig. 9. Interestingly, these percentages are much higher than those obtained for low-seismicity areas. For example, in the case of the cell with the munucipality of Balsorano (L'Aquila Province, Abruzzo Region), where the strongest event occurred on 2019 November 7 with magnitude M_L 4.4 (M_w 4.4), we found that the frequency is below 0.2 per cent.

5 EVALUATING OEF-ITALY THROUGH CONTINGENCY VARIABLES

Here we translate forecasts in predictions defining a threshold for the probabilities, p^*, such that the bins with a higher value of OEF-Italy probability are translated in alarms. The setup of the thresholds p^* automatically induces the identification of four classes of variables in the forecasting experiment:

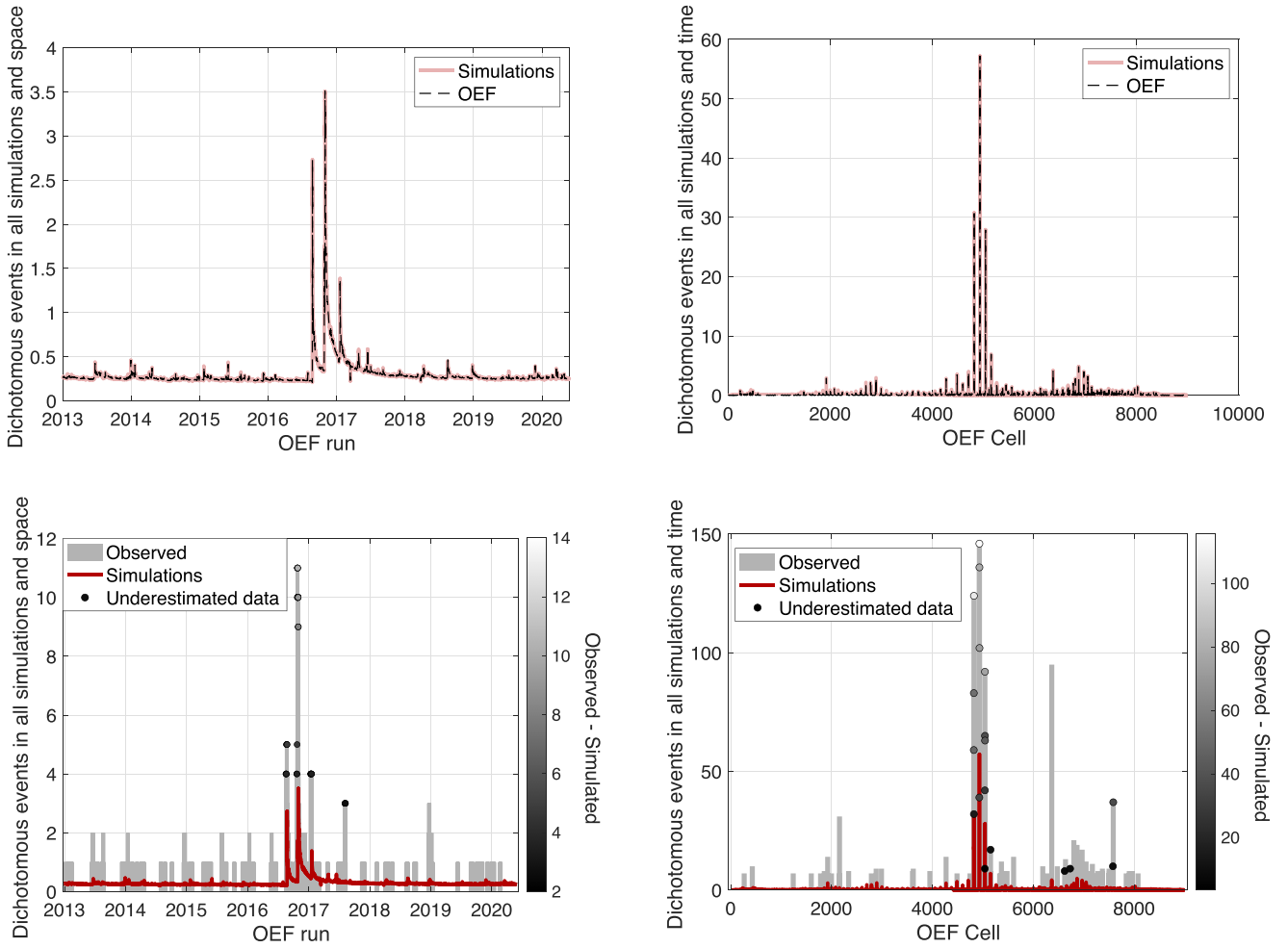

Figure 7. Top panels, left (right): in red, temporal (spatial) trend of the number of OEF forecasts (OEF cells) interested by at least one target event, among the 10 000 synthetic catalogues; in dashed black, sum over the relative ensemble probabilities. Bottom panels, left (right): same simulations of the top panels, but compared to the target events observed in time (space); the circles represent the OEF forecasts (relative OEF cells) corresponding to the forecasts outside the 99 per cent CI obtained from simulations; their colour scales from black to white as the difference between real and synthetic data increases. They are a total of 63 (see Table S2 of the Supporting Information), despite this is not clearly visible from the figure, due to the superimposition of circles corresponding to the outliers that are very close in time, or that occur in the same cell.

(i) True Positives (TP): alarmed spatio-temporal bins (i.e. with $P_{ij} > p^*$), where at least one target event has been observed; SUCCESS.
(ii) False Positives (FP): alarmed spatio-temporal bins (i.e. with $P_{ij} > p^*$), where no target event has been observed; FAILURE.
(iii) True Negatives (TN): non-alarmed spatio-temporal bins (i.e. with $P_{ij} \leq p^*$), where no target event has been observed; SUCCESS.
(iv) False Negatives (FN): non-alarmed spatio-temporal bins (i.e. with $P_{ij} \leq p^*$), where at least one target event has been observed; FAILURE.

These classes are naturally represented by the contingency Table S3 in the Supporting Information, that is the starting point of an analysis based on collected real data. In fact, the four classes can be differently combined to obtain several verification indices, each focusing on a different aspect of the model's forecasting skill.

The four classes can be combined in various ways to build different variables and indices, each one focusing on a different aspect of the experiment. Common denominator is the fact that we are analysing here model's rate values and recorded real earthquakes.

The results are evaluated as a function of the threshold p^*, and this could help the authorities responsible for activating state of alert procedures (Jordan *et al.* 2014). In Fig. 11, we show the percentages of observed target events and spatio-temporal alarmed bins, as well as the latters' rates, for varying p^*. The figure shows that for a threshold $p^* = 10^{-4}$, about 5 per cent of the spatio-temporal bins are alarmed, and their rates sum up to the 60 per cent of the total; in correspondence, about 80 per cent of target events has been observed in those alarmed bins. We stress that the value of $p^* = 10^{-4}$ is just used here to explain the plot, and it is not a 'decisional' threshold. We wish to clarify that the selection of such a threshold cannot be a scientists' task; it is strongly related to several external factors, linked for example to socioeconomic and political aspects, which go beyond scientists' knowledge and authority.

The four classes of contingency variables obtained from the data collected by OEF-Italy as a function of p^* are represented in Fig. 12. The True Negatives give the highest proportion of observations, and in fact the earthquake phenomenon is characterized by a very low frequency of high seismicity, which indeed is ascribed to the extreme (rare) events probability theory. In the OEF-Italy system, the minimum p^* for having at least one target event is 3.43e−6, to which correspond 1791 True Positives, 71 84 120 True Negatives, 234 53 239 False Positives (and 1 False Negative).

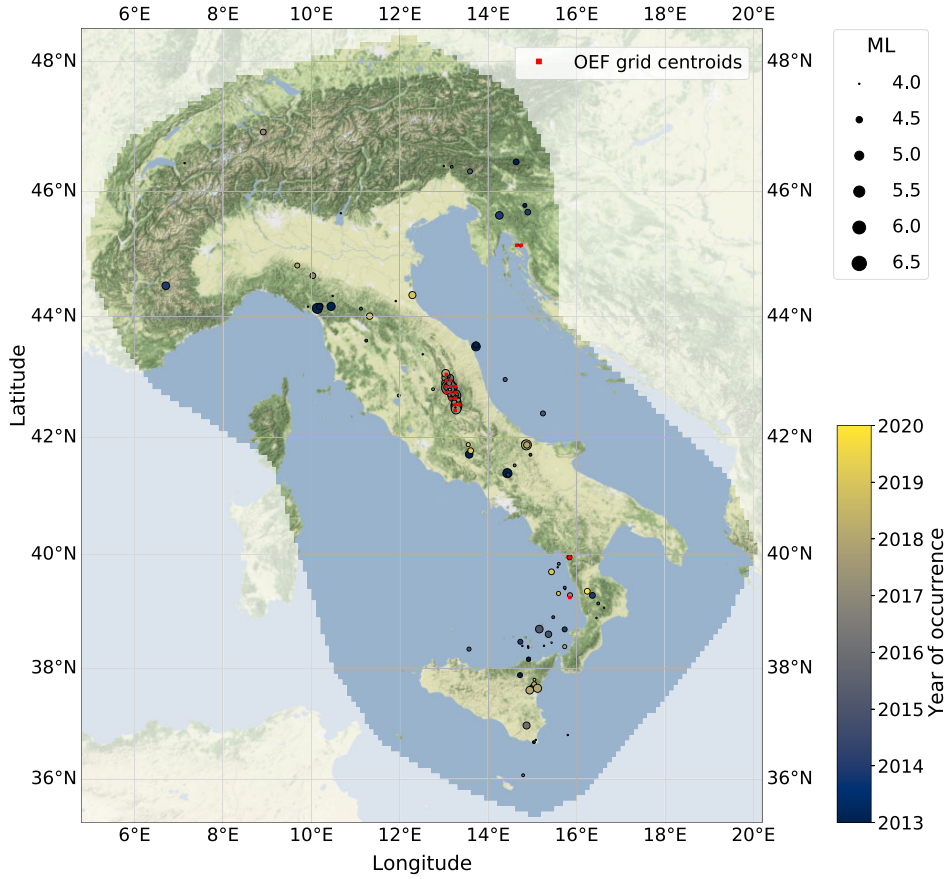

Figure 8. Map of the OEF cells for which the number of forecasts with at least one target event is significantly higher (outside the 99 per cent CI) than expected from the 10 000 simulations. Real seismicity is also shown with size and colour varying with magnitude and temporal occurrence, respectively.

In the following subsections, we will consider the Molchan diagram and the Verification measures to test the OEF-Italy reliability by means of the contingency variables (see Table S1 in the Supporting Information).

5.1 Molchan diagram

The Molchan diagram is used to compare an alarm-based model with a reference model of seismicity defined on the same spatial grid (Molchan 1991; Molchan & Kagan 1992). It is closely related to the Relative Operating Characteristic (ROC) curve (Fawcett 2006), but it has the big advantage of accounting for spatial clustering (e.g. see Zechar & Jordan 2008; Chan *et al.* 2010; Zechar 2010; Shebalin *et al.* 2014). For most space–time predictions, the appropriate reference model is based on previous seismicity, so as to reflect the hypothesis that future earthquakes will occur most likely where they occurred in the past.

This diagram essentially plots the miss rate ν, that is the proportion of target events outside the alarmed area, versus the fraction τ of space–time volume occupied by alarm. More precisely, for each threshold p^*, the pair $(\tau(p^*), \nu(p^*))$ is given by:

$$\tau(p^*) = \{\text{spatio-temporal bins with } P_{ij} > p^*\} = \frac{\text{TP+FP}}{\text{TP+FP+TN+FN}},$$

$$\nu(p^*) = \{\text{spatio-temporal bins with } P_{ij} \leq p^* \text{ but } O_{ij} = 1\} = \frac{\text{FN}}{\text{TP+FN}}.$$

The plot of all these pairs as a function of the threshold p^* results in the Molchan trajectory, which varies between the two estremal values: ($\tau = 1$ and $\nu = 0$), that corresponds to $p^* < \min P_{ij}$, when all the spatio-temporal bins are alarmed and no target event is missed; ($\tau = 0$ and $\nu = 1$), that corresponds to $p^* > \max P_{ij}$, when no spatio-temporal bin is alarmed and every target event is missed. When the diagram shows a diagonal line, no correlation exists between forecasts and observed seismicity: the alarm function (here, the OEF probabilities) does not reflect the distribution of target events and therefore has no predictive skill. An upward/downward arc suggests instead a negative/positive correlation. The ideal result is given by the lowest pair (τ, ν).

For our case of study, the Molchan diagram is shown in Fig. 13. The curve is well below the bisector, highlighting a positive correlation between OEF-Italy forecasts and observed seismicity. Almost all the forecasted earthquakes have location within 75 per cent of the study area with the highest seismicity rate. More in detail, 5 per cent of False Negatives (miss rates, i.e. 95 per cent of True Positives) is located within about the 24 per cent of the alarmed spatio-temporal bins, in correspondence of the threshold $p^* = 2.8\text{e}{-5}$. The contingency table at this threshold value gives 1702 True Positives, 232 40 234 True Negatives, 73 97 125 False Positives and 90 False Negatives. The events that caused these False Negatives are 15, they occurred in 13 different cells, and are situated mostly at the boarders of the Italian territory or off-shore, as can be seen in Fig. 14. The probability gain with respect to the random case can also be appreciated. This represents a good result for the OEF-Italy performance.

We finally compute the Area Skill Score (*ASS*) to understand how well the alarm function is able to estimate the distribution of

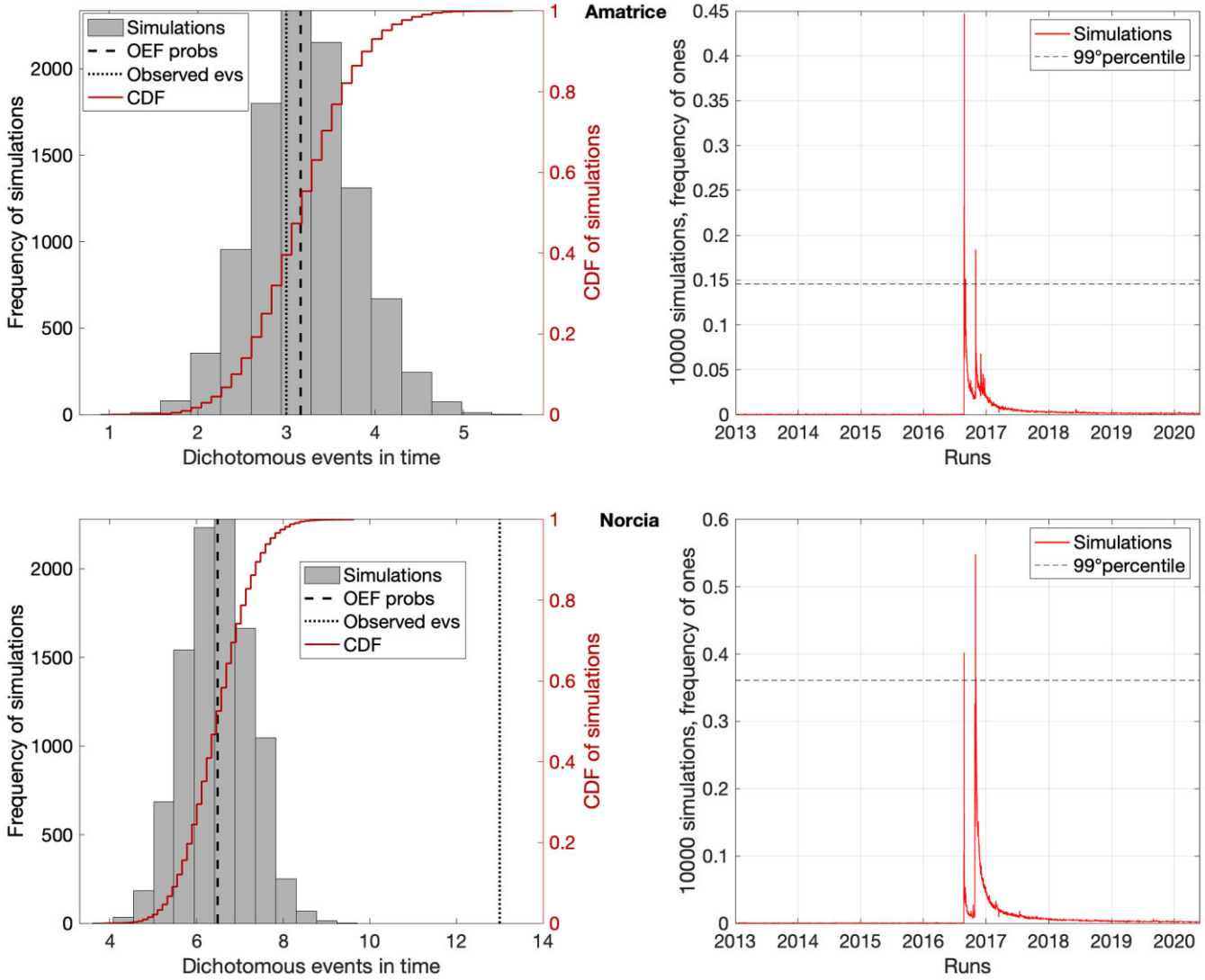

Figure 9. Synthetic analysis focused on two OEF cells, containing the municipalities of Amatrice and Norcia (top and bottom line panels, respectively). The left-hand column contains the histograms (grey bars) and the CDFs (red step lines) of the number of temporal OEF bins with at least one target event, among 10 000 synthetic catalogues. The dashed black lines represent the sum of the overall ensemble probabilities within the specific cell, while the dotted black ones represent the number of temporal bins with at least one observed target event. The right-hand column contains the temporal trend of the number of positive observations among the 10 000 simulations (in red) and the 99th percentile (horizontal dashed black line).

target earthquakes (Zechar & Jordan 2008, 2010). This measure of performance is obtained as the integral of the success rate function, normalized to the space–time volume occupied by alarm:

$$ASS(\tau) = \frac{1}{\tau}\int_0^{\tau} [1 - \nu(t)]\, dt;$$

the larger the statistic, the better the performance. The *ASS* for the OEF-Italy system is given in Fig. 15. For the largest τ, we obtain $ASS \sim 0.7$.

5.2 Verification measures

The verification measures (Jolliffe & Stephenson 2011) represent a family of indices obtained as a function of the forecasts, the observations and/or their relationship. They could be very useful to understand and evaluate the forecasting skill of any binary system, like the OEF-Italy one, that in fact can be formalized in terms of the dichotomous variables '0' or '≥1' expected VS observed target earthquakes. They are typically adopted in weather forecasting applications but, since they are robust general methodologies able to diagnose the performance when forecasts and observations are both available on the same spatial grid (Casati *et al.* 2008; Gilleland *et al.* 2009, 2010, see also https://www.cawcr.gov.au/projects/verification/), we believe they can conform to our case, as well. Some authors have already used some verification methods for earthquake forecasting (Holliday *et al.* 2005; Chen *et al.* 2006; Murru *et al.* 2009), but we stress that much attention has to be paid on the proper choice of indices, since earthquakes cluster in space and time. For example, the ROC method very often used for dichotomous forecasts assumes that the events are equally likely to occur anywhere in space, therefore it represents a weak tool when evaluating earthquake forecasts (Zechar 2010).

The proper choice has to be done by considering the specific context of analysis, which in our case consists in dichotomous variables and rare events forecasting.

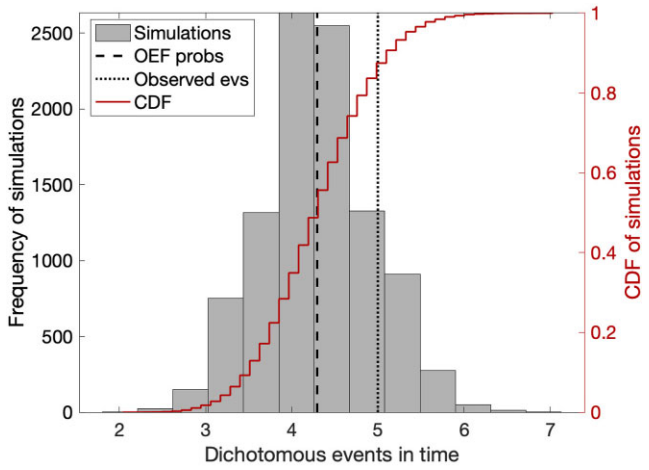

Figure 10. The same as the bottom left panel in Fig. 9, but removing from the analysis 1 d after each M_L 5.4+ event (2 events in total).

In what follows, we will consider two main subfamilies: the descriptive and the performance measures. The first allows us to make statistical inference on the system involved, while the second are effectively used to evaluate the performance of its forecasting skill.

5.2.1 Descriptive measures

The descriptive measures are obtained as a function of forecasts and observations, but not of their correspondence. They cannot be used to evaluate the performance of the system, but these variables are basic descriptive statistics that can help to infer underlying properties of the system itself.

The Base Rate (BR) is the percentage of positive observations, that is, a sample estimate of the unconditional marginal occurrence probability of the observed events. Although it is not a characteristic of forecasting skill, many performance measures are related to it, and therefore are sensitive to variations due to the natural variability of the observed data (Jolliffe & Stephenson 2011). The Rate of Alarms (RA) is instead a sample estimate of the marginal probability of a forecast of occurrence. It is indeed the τ variable in the Molchan diagram. The ratio between the two descriptive measures just described gives the frequency bias, or Index of Distorsion (ID). Since it is computed as the number of forecasts of occurrence over the number of actual occurrences, an index ID = 1 results in the unbiased forecasts: the skill is perfect when BR = RA, that is, no false outcome. Instead, ID > 1 (ID < 1) indicates an overestimation (underestimation) of the positive observations. The explicit formulations of the three above verification measures are given in eq. (2), while Fig. 16 shows the relative trends obtained in the case of the OEF-Italy system, for varying p^*. The plot points out that the perfect skill is obtained for $p^* = 0.0386$. Finally, the BR value constantly equal to 5.85e−5 is compatible to the fact that we are forecasting (extreme) rare events.

$$\begin{aligned} \mathrm{BR} &= \frac{\mathrm{TP+FN}}{\mathrm{TP+FP+TN+FN}}, \\ \mathrm{RA} &= \frac{\mathrm{TP+FP}}{\mathrm{TP+FP+TN+FN}} = \tau, \\ \mathrm{ID} &= \frac{\mathrm{TP+FP}}{\mathrm{TP+FN}} = \frac{\mathrm{RA}}{\mathrm{BR}}. \end{aligned} \tag{2}$$

5.2.2 Performance measures

The performance measures constitute a subset of the verification measures that focuses on the correspondence between forecasts and observations. They are a very powerful tool to evaluate the reliability of a binary-based forecasting system such as OEF-Italy.

The first performance measure that we discuss is the Probability of Detection (POD), that is, the proportion of occurrences correctly forecasted. It is a sample estimate of the probability of the event to be forecasted, conditional to the fact that this event has been observed:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}.$$

This measure is also know as the hit rate, in fact it is the complementary of the miss rate ν in the Molchan diagram. A threshold probability of 0 (of 1) means that the occurrence is always (never) forecasted and then POD = 1 (POD = 0).

Since the forecasting skill depends on the best trade-off between maximizing the number of hits and minimizing the number of false alarms, the POD alone is not sufficient for measuring the forecasting performance of the system. We therefore introduce the False Alarm Ratio (FAR): a sample estimate of the probability of a false alarm, conditional to the fact that the occurrence has been forecasted:

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP+FP}}.$$

The reliability of this performance measure is strictly connected to its dependence on both the Base Rate and the optimal threshold p^*: it equals 0 when the skill is perfect, while for zero skill it is FAR = 1 − BR. We stress also that ID = $\frac{\mathrm{POD}}{1-\mathrm{FAR}}$, therefore when POD > 1 − FAR (POD < 1 − FAR), the binary system of forecasts tends to overestimate (underestimate) the positive observations.

An additional measure very often quoted with POD and FAR and used in the literature (e.g. see Golian *et al.* 2011) is the Critical Success Index (CSI). It still can be calculated without using the frequency of correct rejections, property which makes the CSI a very useful measure for evaluating the forecasts of rare events (as in OEF). It is defined as

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}},$$

and it can be regarded as a sample estimate of the probability of a hit, conditional to the fact that this event has been either forecasted, or observed, or both. In case of a perfect skill, max CSI = 1; instead, the minimum value min CSI = 0 is got when there are no True Positives. The values of CSI corresponding to zero skill could be obtained at any point of the interval [0, BR], depending on the proportion of forecasts of occurrence to non-occurrence in the sample. This performance measure strongly depends on the probability threshold p^*, whose optimal value is obtained for $p^* = \frac{\mathrm{CSI}}{1+\mathrm{CSI}}$.

Since POD, FAR, ID and CSI are key quality measures of a dichotomous system forecast, they are often represented in a single 'verification diagram' (Roebber 2009), which allows us to give an immediate visualization of the performance skill. The idea follows from Taylor (2001) and Lambert & Boer (2001), both exploiting a way to represent the geometric relationship between the three measures: 'correlation', 'normalized root-mean-square difference (RMSd)' and 'variance' of model performance. More precisely, a good model is such that forecast and observation simultaneously have high correlation, small RMSd and similar variances. Similarly, the performance skill of a dichotomous system is good when POD and CSI are high (∼1), FAR is small (∼0, or SR=1-FAR ∼1) and BR is similar to RA (i.e. the percentage of positive

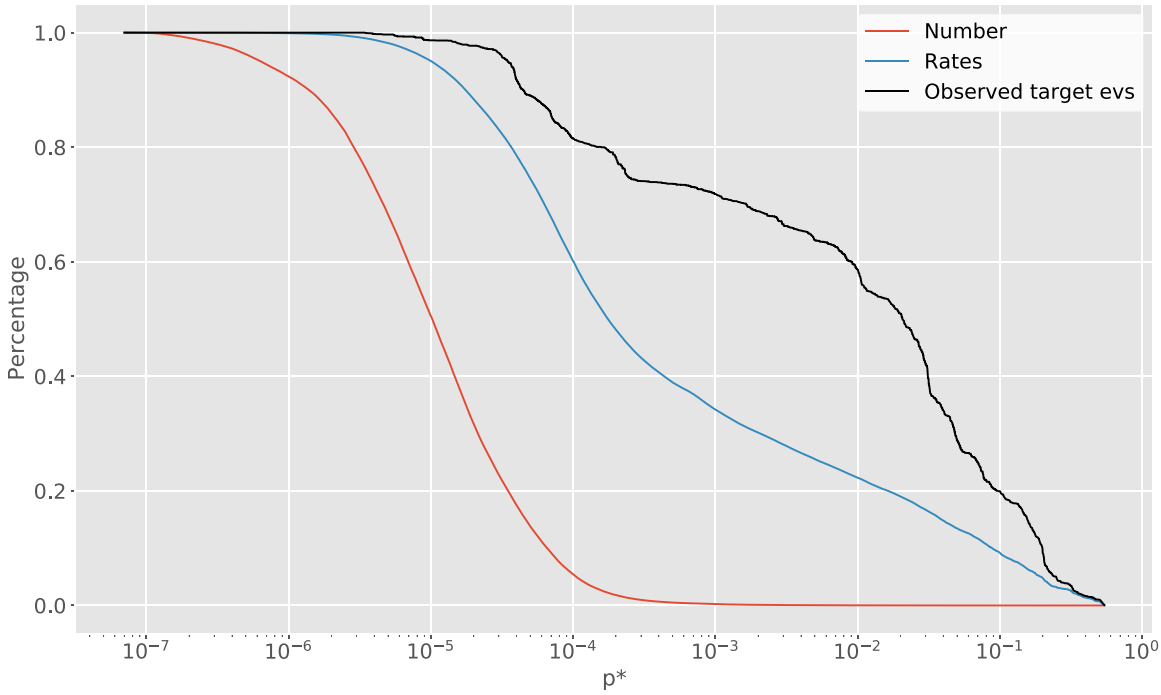

Figure 11. Percentage of alarmed spatio-temporal bins with respect to the threshold probabilities p^*. The red line represents the percentage number of alarmed bins, while the blue one is the percentage sum of their rates. The percentage of observed target events is also shown as a black line.

Figure 12. Four classes of variables of the contingency table (see Table S3 in the Supporting Information), obtained from the data collected by OEF-Italy, as a function of the threshold p^*.

observation approaches the rate of alarms: ID ∼1). These are indeed the quantities represented in the verification diagram which, in the case of the OEF-Italy system, is given in Fig. 17. To obtain this diagram we used the R Verification Package available at https://CRAN.R-project.org/package=verification, and we considered a subset of probability thresholds to reduce computational cost. Sampling uncertainty is automatically computed by the software, and is represented in the plot as crosshairs. Moving along the bisector towards the upper right of the diagram indicates an increase in absolute accuracy. As expected, in our case we observe again that the best trade-off (highest POD, SR and CSI, ID∼1) is obtained for $p^* \sim 0.038$, indicated with a red circle in the figure. Still, several

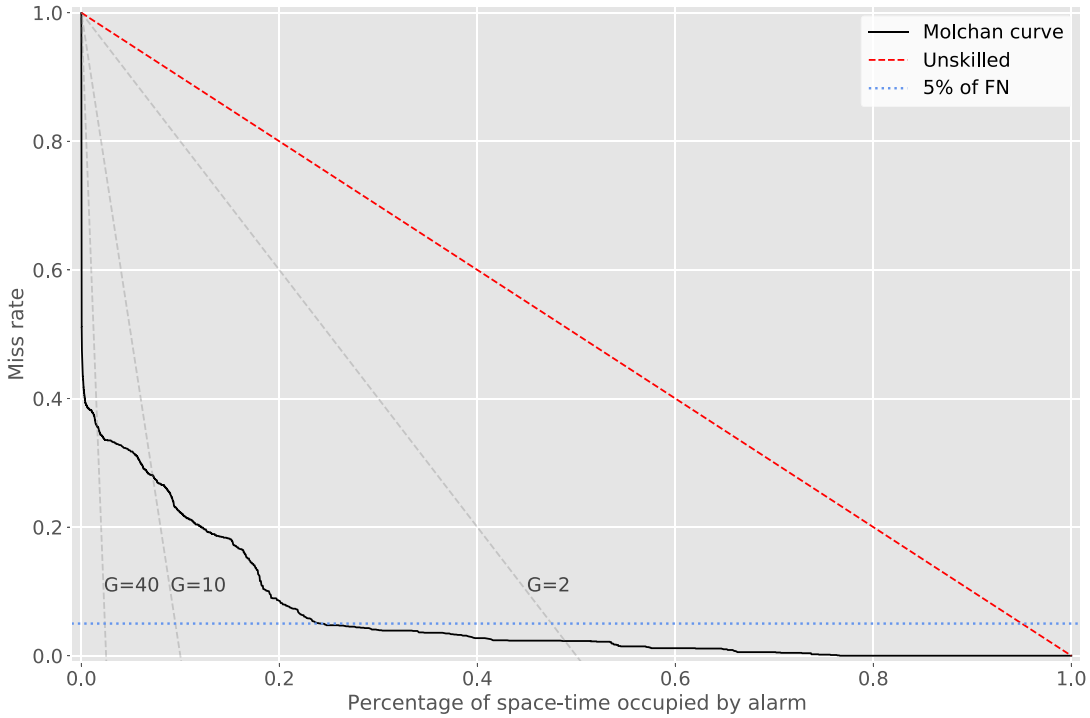

Figure 13. Molchan diagram obtained from the data collected by OEF-Italy. The red dashed line corresponds to unskilled forecasts; the blue dotted line identifies the 5 per cent of Miss rate (False Negatives); the grey dashed lines indicate the probability gain $G = \frac{1-\nu}{\tau}$.

authors claim that the ideal level of statistical performance is obtained in the area delimited by a CSI higher than 0.6 (e.g. Roberts *et al.* 2012), and this highlights a large room for improvement that the OEF-Italy system might have. Finally, we stress that if we exclude 1 d after the 6 strongest M_L 5.4+ events, we obtain a diagram showing a worse accuracy: this is what expected, as in this way we remove True Positives cases.

It is important to say that 2×2 contingency tables, like in Table S3 of the Supporting Information, have three degrees of freedoms, and can be completely expressed by the three verification measures POD, ID and POFD$= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$ (Stephenson 2000), where the latter index is the so-called false alarms rate, also known as Probability of False Detection, and it indicates the proportion of non-occurrences that were incorrectly forecasted. In this sense, the diagram of Fig. 17 is incomplete as it neglects the number of times when a null event is forecasted but no observation is recorded (True Negatives). However, in rare event forecasting like our case, neglecting this term can be even useful, as the verification of plenty trivial non-events can inflate the skill (Roebber 2009).

The last score we discuss is the Extremal Dependence Index (EDI, Ferro & Stephenson 2011). This is independent of BR, and precisely a function of only POD and POFD:

$$\mathrm{EDI} = \frac{\ln \mathrm{POFD} - \ln \mathrm{POD}}{\ln \mathrm{POFD} + \ln \mathrm{POD}},$$

It is specific to measure the correlation between forecasts and observations in case of rare events, and ranges in $[-1, 1]$: the value EDI $= 1$ indicates a perfect positive correlation, and is obtained when the False Negatives are 0 (the other contingency variables being non-zero). Therefore, EDI can be optimized for biased forecasts. The trend of this measure for the OEF-Italy system is given in Fig. 18 as a function of p^*. We observe that it is almost completely included in the range [0.6, 1], thus indicating a weak-to-strong positive correlation between forecasts and observations. This confirms the results obtained by the Molchan diagram.

5.3 Reliability diagram

The final analysis we consider to assess the forecasting skill of the OEF-Italy system is the reliability diagram (Jolliffe & Stephenson 2011; Bröcker & Smith 2007). It is a diagnostic check for the consistency of probability forecasts of dichotomous events with respect to the relative observed frequencies. More precisely, it is the plot of the expected cumulative distribution of forecast values, and the observed cumulative proportion of observations, which shows how much the frequency of any dichotomous event is consistent with the relative probability forecast.

The procedure requires the forecasts to be grouped in a countable number of representative bins, whose definition is completely arbitrary. In our case, in order to avert any loss of information about the forecasting model's performance, we use exactly the spatio-temporal bins produced by OEF. Given the ensemble probabilities P_{ij}, for $(i, j) \in N_t \times N_c$, expected and observed CDFs are therefore obtained as:

$$F_{\mathrm{for}}(p^*) = \frac{\sum_{(i,j)\in I_{p^*}} P_{ij}}{\sum_{(i,j)\in N_t \times N_c} P_{ij}}, \qquad F_{\mathrm{obs}}(p^*) = \frac{\sum_{(i,j)\in I_{p^*}} O_{ij}}{\sum_{(i,j)\in N_t \times N_c} O_{ij}},$$

where $I_{p^*} = \{(i, j) \in N_t \times N_c \mid P_{ij} \leq p^*\}$.

Fig. 19 shows the reliability diagrams obtained for the OEF-Italy system considering all the data (top panel), removing 1 d after the 6 events with M_L 5.4+ (middle panel), and excluding the forecasts relative to the entire temporal interval of the Central Italy sequence (bottom panel). The cumulative proportion of earthquakes (blue lines) fit well the expected CDFs (red lines) only in the period preceeding the Central Italy sequence, or excluding that data. The

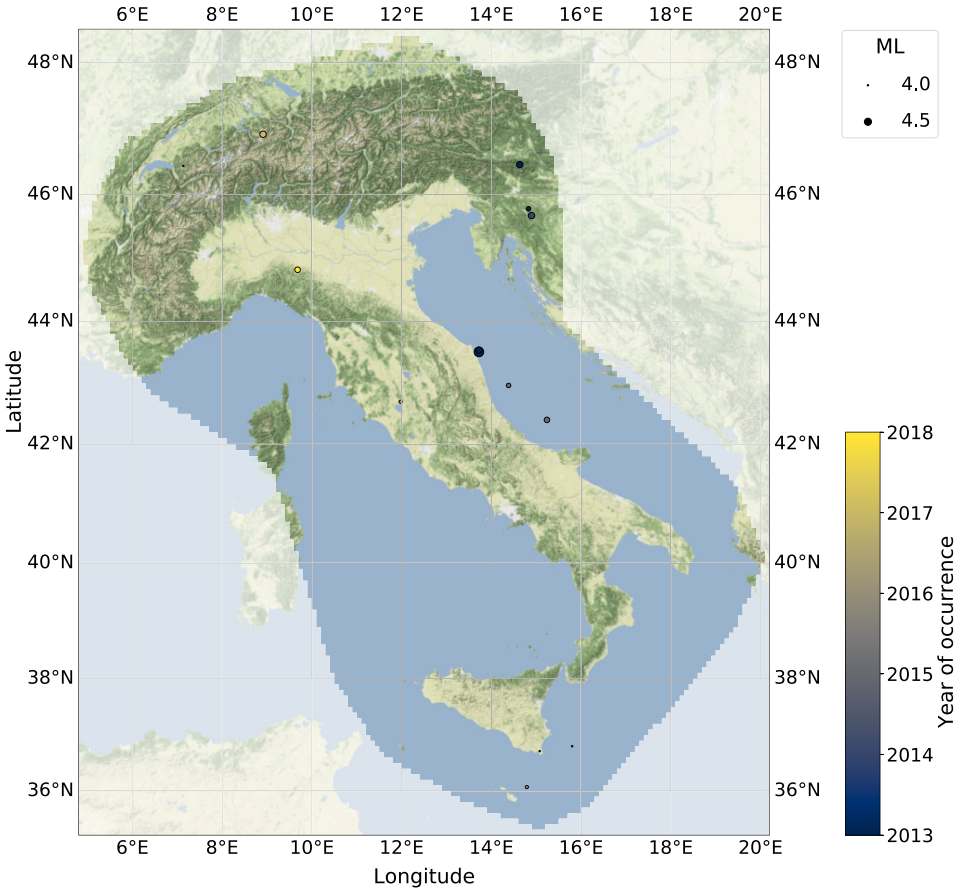

Figure 14. Seismic map of the 15 earthquakes (occurred in 13 different cells) that caused the 90 False Negatives identified in the Molchan diagram with a threshold $p^* = 2.8e{-}5$, to which corresponds about the 24 per cent of the alarmed spatio-temporal bins (see also Fig. 13).

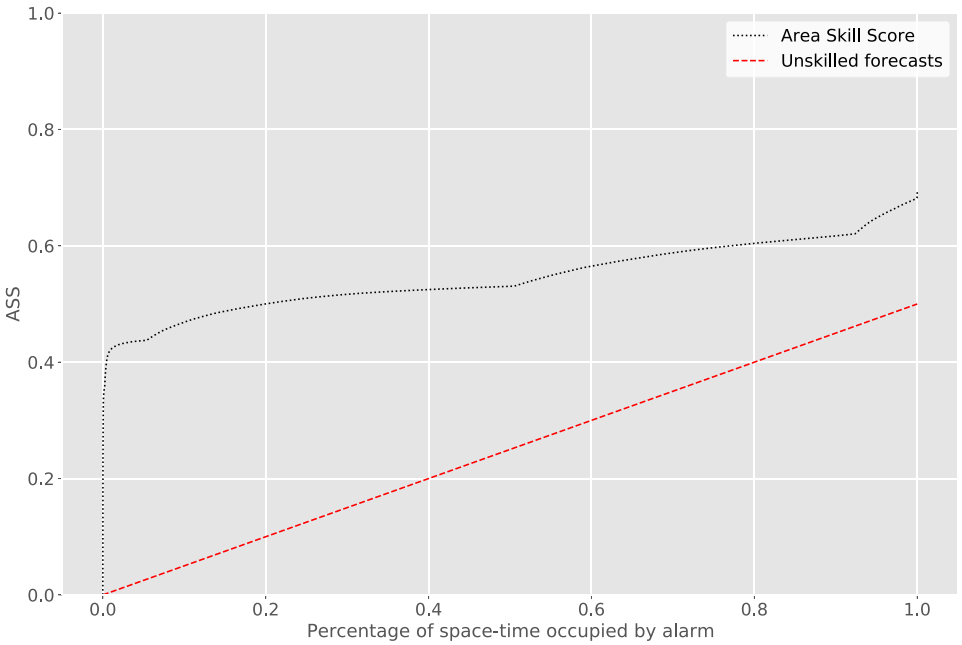

Figure 15. ASS diagram obtained from the data collected by OEF-Italy. The red dashed line corresponds to unskilled forecasts.

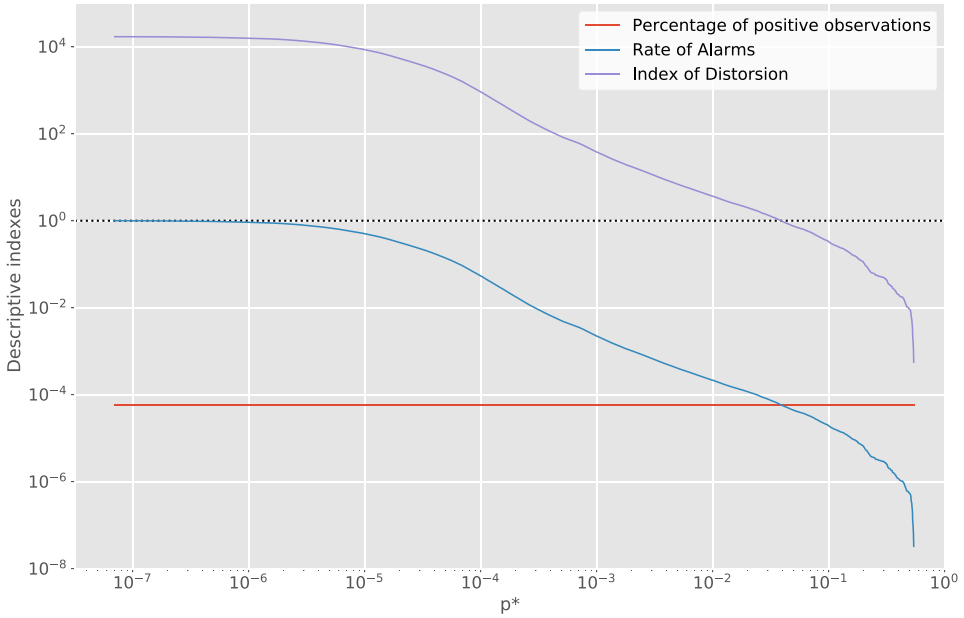

Figure 16. BR, RA and ID obtained for the OEF-Italy system, for varying p^*.

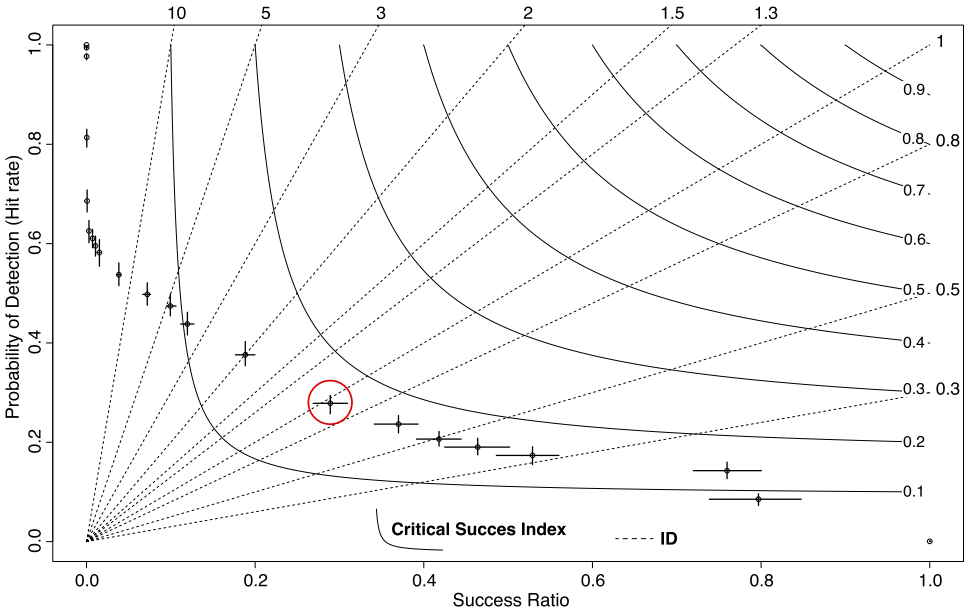

Figure 17. Verification diagram obtained for the OEF-Italy system, for a subset of probability thresholds p^*. As specified in the plot, contour (dotted) lines indicate CSI (ID). Crosshairs represent sampling uncertainty. Finally, the red circle encloses the probability threshold which gives the highest POD, SR and CSI (best absolute accuracy of the system).

OEF-Italy model is shown to overestimate the observed cumulative proportion of earthquakes from the threshold 10^{-4} on, and this is due again to the Central Italy sequence. In fact, the discrepancy is barely reduced when removing the 6 d after the strongest events, and almost disappears by excluding the entire sequence, case in which a good agreement between the two CDFs is obtained. We finally observe that the increasing velocity of the observed seismicity is slower than that of the expected seismicity in the range $p^* \in [10^{-4}, 10^{-2}]$, while it is faster thereafter.

6 CONCLUSIONS

The main aim of this work was to evaluate the reliability of the short-term seismic forecasts produced by the OEF-Italy system during its first years of real-time operativity, in comparison with the real earthquake catalogue recorded in the same period. We consider as target all the events with depth $\leq$30 km and local magnitude $M_L \geq 4.0$, occurred in the grid covering the whole Italian territory represented in Fig. 1, within the testing temporal interval 2013 January 01–2020 May 26. The probabilistic forecasts were delivered

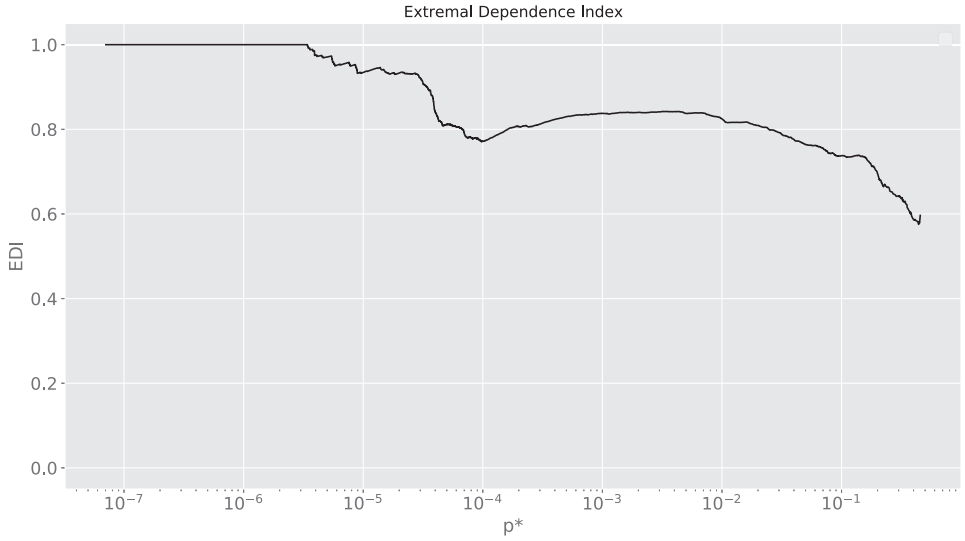

Figure 18. EDI obtained for the OEF-Italy system, for varying p^*.

by the OEF-Italy system at the midnight of each day and after the occurrence of any $M_L \geq 3.5$ earthquake, and they consist in the weekly forecasts of the target events from the ensemble model, that is a weighted combination of three versions of the models ETAS, ETES and STEP, largely used in statistical seismology.

For a proper interpretation of the OEF forecasts, it is worth to specify that the models included in the system need to be quickly updated to integrate the effect of any earthquake occurring during any sequence. This may be challenging immediately after a large shock, which indeed may produce an abundance of triggered earthquakes that are largely incomplete in real-time earthquake catalogues. Unavoidably, this leads to the underestimation of the earthquake activity in the aftermath of a large event.

To assess the reliability of such probabilistic forecasts, in this paper we accounted for both synthetic and real dichotomous observations in each of the $N_t \times N_c$ spatio-temporal bins considered in the analysis, where N_t is the number of forecasts produced by the OEF-Italy system in the testing time window, and N_c is the number of the spatial grid's cells over which the analysis is performed.

All the statistical methodologies applied for evaluating the performance of the OEF-Italy system show an overall good agreement between expected and observed seismicity, with one exception related to the Central Italy sequence. The discrepancy in this case is mainly due to the strong incompleteness introduced by the sequence, as well as to the nature of the probabilistic models currently involved in OEF, which feed on progressively occurring events and need time to grasp how much productive is going to be any starting aftershock sequence. Another important aspect to specify is that, during relevant sequences, the minimum reporting threshold to be recorded in real-time in the INGV Seismic Monitoring room in Rome is raised to M_L 4.0. Only a revision at a later stage will include smaller events in the database. Therefore, it could happen that an event with M_L in [3.5, 4) occurs, is not recorded at a first stage and therefore the system does not produce the run, but this event is included in the catalogue after a certain period of revision. This may influence any retrospective analysis of the forecasting skill of the system.

The results we obtained highlight the potential of the OEF probabilistic forecasting experiment for the short-term earthquake prediction in Italy. At the same time, they show some rooms for improvement on two major fronts. First, the OEF-Italy system still needs some hand corrections during the first hours of an energetic seismic sequence to take into account the strong catalogue incompleteness, which causes a severe underestimation of the expected seismicity. Second, the inclusion of additional models, preferably based on different assumptions, could give a relevant additional contribution with respect to the clustering models used so far. Besides that, it is necessary to overcome the technical problems related to the computational time. To date, the system checks the earthquake catalogue to identify M_L 3.5+ events every 15 min. This means that if the last check occurred at 00:00 and a strong event occurs at 00:01, it will be recorded at 00:15, thus introducing a temporal delay that could entail underestimation.

We are now working to make several adjustments to the system in the near future. In order to account for STAI, the RESTORE algorithm by Stallone & Falcone (2021) will be included in OEF-Italy: it implements a stochastic gap-filling method that detects STAI gaps and reconstructs the missing events in a space–time–magnitude domain, thus extending the work by Zhuang *et al.* (2017, 2020), that replenish the portions of an incomplete seismic catalogue through empirical functions describing only the time–magnitude range of missing data.

The possibility to estimate in OEF-Italy the model's parameters by means of a Bayesian procedure, as proposed in Omi *et al.* (2014), is also being discussed to reduce uncertainty in the forecasts (Michael *et al.* 2020; van der Elst *et al.* 2022). During the sequences of L'Aquila 2009 (Chiaraluce *et al.* 2011) and Pianura Padana Emiliana 2012 (Scognamiglio *et al.* 2012), we made a first attempt of a daily calibration of the OEF-Italy models. However, in both those cases, we observed an overestimation of the events' number in the tails of the sequences. This is likely due to the fact that the parameters were estimated over a considerably large amount of data, thus making the estimation so firmly stable that the model's temporal decay was lower than the effective course of the sequence. A region-specific parameter estimation could be a first step in the OEF-Italy system parameters' updating. A particular attention is needed to address the issue of calibrating the OEF-Italy models in real-time, also because this would imply a remarkable computational cost. Because of these reasons, we are willing to open a

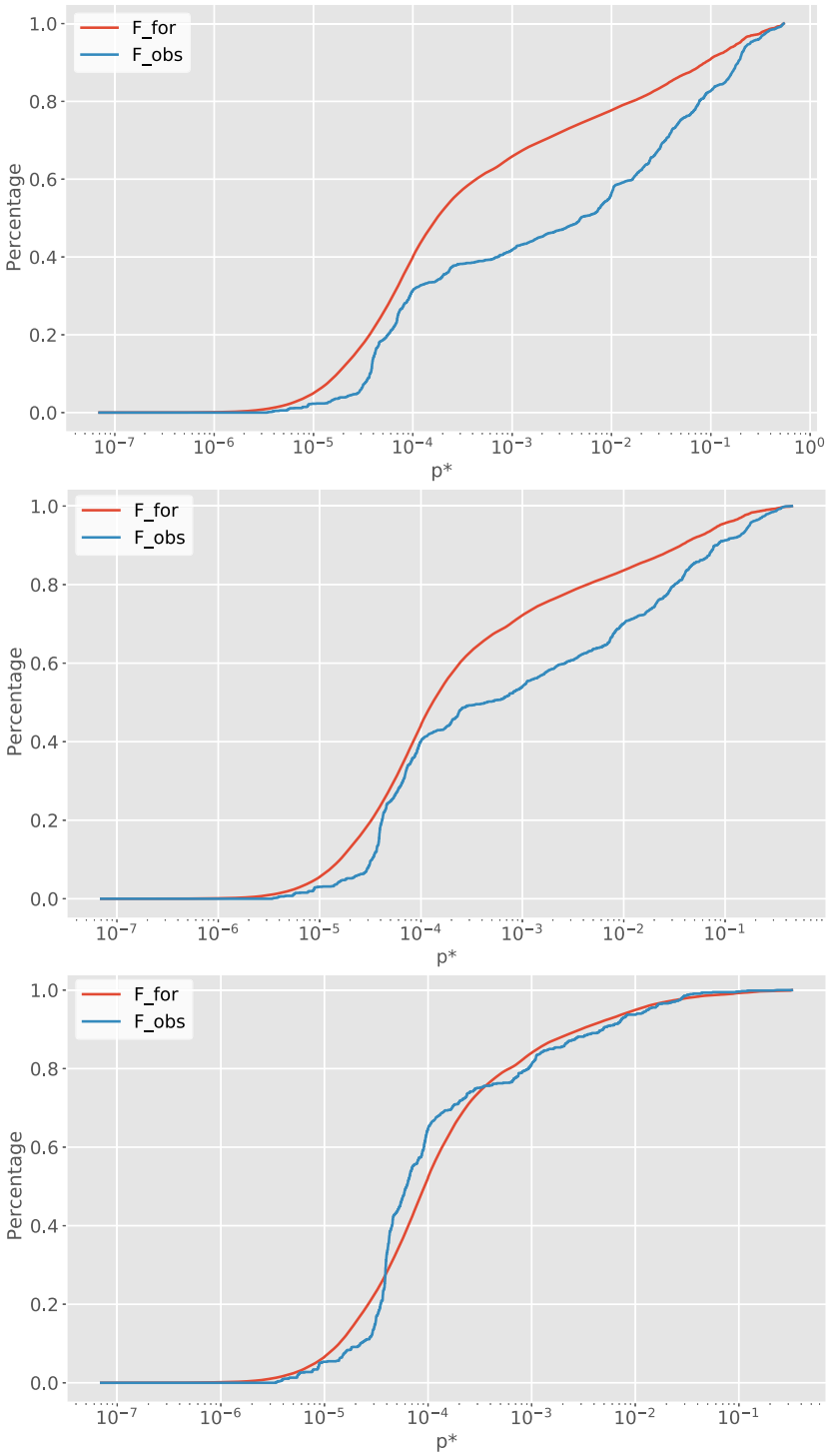

Figure 19. Reliability diagrams comparing the CDFs of OEF-expected (red) and observed (blue) seismicity. Top, middle and bottom panels are respectively obtained from all the data, removing 1 d after the 6 events with M_L 5.4+, and excluding the forecasts relative to the temporal interval of the Central Italy sequence.

discussion with experts in the field, to try to find the best solution for the Italian system.

Eventually, we are also planning to include different models in the OEF-Italy system, both explicitly accounting for incompleteness as in Mizrahi *et al.* (2021), and more physics-based models (Mancini *et al.* 2019). We strongly believe that all these improvements will help to increase the OEF-Italy reliability and reveal how this experiment for short-term seismic prediction in Italy is even more useful, useable and has a greater potential than it is already believed to have.

ACKNOWLEDGMENTS

This study has benefited from funding provided by the Italian Presidenza del Consiglio dei Ministri—Dipartimento della Protezione Civile (DPC) to the Seismic Hazard Center (Centro di Pericolosità Sismica, CPS, at the Istituto Nazionale di Geofisica e Vulcanologia, INGV). This paper does not represent DPC official opinion and policies. The authors thank the CPS coordinator André Herrero for useful discussion, and Anna Maria Lombardi for her considerable help in the initial stage of this study. The authors are also grateful to the Editor Margarita Segou, the reviewer Andy Michael and the other anonymous reviewer for their useful comments and suggestions, that helped improving the quality of the manuscript (Omori 1894; Gutenberg & Richter 1944; Ogata 1988, 1989; Kagan 1991; Vere-Jones 1995; Frankel 1995; Ogata 1998; Wiemer & Katsumata 1999; Wiemer *et al.* 2002; Marzocchi *et al.* 2020).

DATA AVAILABILITY

The data underlying this paper will be shared on reasonable request to the corresponding author.

REFERENCES

Bayona, J.A., Savran, W.H., Rhoades, D.A. & Werner, M.J., 2022. Prospective evaluation of multiplicative hybrid earthquake forecasting models in California, *Geophys. J. Int.,* **229**(3), 1736–1753.

Becker, J.S., Potter, S.H., McBride, S.K., Doyle, E.E., Gerstenberger, M.C. & Christophersen, A., 2020. Forecasting for a fractured land: A case study of the communication and use of aftershock forecasts from the 2016 Mw 7.8 Kaikōura earthquake in Aotearoa New Zealand, *Seismol. Res. Lett.,* **91**(6), 3343–3357.

Bröcker, J. & Smith, L.A., 2007. Increasing the reliability of reliability diagrams, *Weather Forecast.,* **22**(3), 651–661.

Casati, B., *et al.*, 2008. Forecast verification: current status and future directions, *Meteorol. Appl.,* **15**(1), 3–18.

Chan, C.-H., Sørensen, M., Dietrich, S., Grünthal, G., Heidbach, O., Hakimhashemi, A. & Cataldi, F., 2010. Forecasting Italian seismicity through a spatio-temporal physical model: importance of considering time-dependency and reliability of the forecast, *Ann. Geophys.,* **53**(3), 129–140.

Chen, C.-C., Rundle, J.B., Hsien-Chi, L., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L. & Tiampo, K.F., 2006. From tornadoes to earthquakes: forecast verification for binary events applied to the 1999 Chi-Chi, Taiwan, earthquake, *TAO: Terrest. Atmos. Oceanic Sci.,* **17**(3), 503–516.

Chiaraluce, L. *et al.*, 2011. The 2009 L'Aquila (central Italy) seismic sequence, *Boll. Geofis. Teorica Ed Appl.,* **52,** 367–387.

Falcone, G., Console, R. & Murru, M., 2010. Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST, *Ann. Geophys.,* **53**(3), 41–50.

Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recog. Lett.,* **27**(8), 861–874.

Ferro, C. & Stephenson, D., 2011. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events, *Weather Forecast.,* **26,** doi:10.1175/WAF-D-10-05030.1.

Frankel, A., 1995. Mapping seismic hazard in the Central and Eastern United States, *Seismol. Res. Lett.,* **66**(4), 8–21.

Gerstenberger, M.C., Wiemer, S., Jones, L.M. & Reasenberg, P.A., 2005. Real-time forecasts of tomorrow's earthquakes in California, *Nature,* **435**(7040), 328–331.

Gilleland, E., Ahijevych, D.A., Brown, B.G., Casati, B. & Ebert, E.E., 2009. Intercomparison of spatial forecast verification methods, *Weather Forecast.,* **24**(5), 1416–1430.

Gilleland, E., Ahijevych, D.A., Brown, B.G. & Ebert, E.E., 2010. Verifying forecasts spatially, *Bull. Am. Meteorol. Soc.,* **91**(10), 1365–1376.

Golian, S., Saghafian, B., Elmi, M. & Maknoon, R., 2011. Probabilistic rainfall thresholds for flood forecasting: evaluating different methodologies for modelling rainfall spatial correlation (or dependence), *Hydrol. Processes,* **25**(13), 2046–2055.

Gutenberg, B. & Richter, C.F., 1944. Frequency of earthquakes in California, *Bull. seism. Soc. Am.,* **34**(8), 185–188.

Helmstetter, A., Kagan, Y.Y. & Jackson, D.D., 2006. Comparison of short-term and time-independent earthquake forecast models for Southern California, *Bull. seism. Soc. Am.,* **96**(1), 90–106.

Holliday, J.R., Nanjo, K.Z., Tiampo, K.F., Rundle, J.B. & Turcotte, D.L., 2005. Earthquake forecasting and its verification, *Nonlinear Process. Geophys.,* **12**(6), 965–977.

Jolliffe, I.T. & Stephenson, D.B., 2011. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* 2nd edn, John Wiley & Sons.

Jordan, T. *et al.*, 2011. Operational earthquake forecasting. state of knowledge and guidelines for utilization, *Ann. Geophys.,* **54**(4) , doi:10.4401/ag-5350.

Jordan, T.H., Marzocchi, W., Michael, A.J. & Gerstenberger, M.C., 2014. Operational earthquake forecasting can enhance earthquake preparedness, *Seismol. Res. Lett.,* **85**(5), 955–959.

Kagan, Y.Y., 1991. Likelihood analysis of earthquake catalogues, *Geophys. J. Int.,* **106**(1), 135–148.

Kagan, Y.Y., 2004. Short-term properties of earthquake catalogs and models of earthquake source, *Bull. seism. Soc. Am.,* **94**(4), 1207–1228.

Lambert, S.J. & Boer, G.J., 2001. CMIP1 evaluation and intercomparison of coupled climate models, *Climate Dynam.,* **17**(2), 83–106.

Lilliefors, H., 1969. On the Kolmogorov-Smirnov test for the Exponential Distribution with mean unknown, *J. Amer. Statis. Assoc.,* **64,** 387–389.

Lippiello, E., Cirillo, A., Godano, C., Papadimitriou, E. & Karakostas, V., 2019. Post seismic catalog incompleteness and aftershock forecasting, *Geosciences,* **9**(8), 355. Available from: https://doi.org/10.3390/geosciences9080355.

Lombardi, A.M. & Marzocchi, M., 2010. The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment, *Ann. Geophys.,* **53**(3), 155–164.

Mancini, S., Segou, M., Werner, M.J. & Cattania, C., 2019. Improving physics-based aftershock forecasts during the 2016–2017 Central Italy Earthquake Cascade, *J. geophys. Res.: Solid Earth,* **124**(8), 8626–8643.

Marzocchi, W., 2012. Putting science on trial, *Phys. World,* **25**(12), 17, doi:10.1088/2058-7058/25/12/27.

Marzocchi, W. & Lombardi, A.M., 2009. Real-time forecasting following a damaging earthquake, *Geophys. Res. Lett.,* **36**(21), L21302, doi:10.1029/2009GL040233.

Marzocchi, W., Zechar, J.D. & Jordan, T.H., 2012. Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. seism. Soc. Am.,* **102**(6), 2574–2584.

Marzocchi, W., Lombardi, A.M. & Casarotti, E., 2014. The establishment of an operational earthquake forecasting system in Italy, *Seismol. Res. Lett.,* **85**(5), 961–969.

Marzocchi, W., Taroni, M. & Falcone, G., 2017. Earthquake forecasting during the complex amatrice-norcia seismic sequence, *Sci. Adv.,* **3,** e1701239, doi:10.1126/sciadv.1701239.

Marzocchi, W., Spassiani, I., Stallone, A. & Taroni, M., 2020. How to be fooled searching for significant variations of the b-value, *Geophys. J. Int.,* **220**(3), 1845–1856.

Michael, A.J., 2012. Do aftershock probabilities decay with time?, *Seismol. Res. Lett.,* **83**(4), 630–632.

Michael, A.J. *et al.*, 2020. Statistical seismology and communication of the USGS Operational Aftershock Forecasts for the 30 November 2018 Mw 7.1 Anchorage, Alaska, Earthquake, *Seismol. Res. Lett.,* **91**(1), 153–173.

Mizrahi, L., Nandan, S. & Wiemer, S., 2021. Embracing data incompleteness for better earthquake forecasting, *J. geophys. Res.: Solid Earth,* **126**(12), e2021JB022379, Available from: https://doi.org/10.1029/2021JB022379.

Molchan, G.M., 1991. Structure of optimal strategies in earthquake prediction, *Tectonophysics,* **193**(4), 267–276.

Molchan, G.M. & Kagan, Y.Y., 1992. Earthquake prediction and its optimization, *J. geophys. Res.: Solid Earth,* **97**(B4), 4823–4838.

Murru, M., Console, R. & Falcone, G., 2009. Real time earthquake forecasting in Italy, *Tectonophysics,* **470**(3–4), 214–223.
Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.,* **83**(401), 9–27.
Ogata, Y., 1989. Statistical model for standard seismicity and detection of anomalies by residual analysis, *Tectonophysics,* **169**(1–3), 159–174.
Ogata, Y., 1998. Space-time point process models for earthquake occurrences, *Ann. I. Stat. Math.,* **50**(2), 379–402.
Omi, T., Ogata, Y., Hirata, Y. & Aihara, K., 2014. Estimating the ETAS model from an early aftershock sequence, *Geophys. Res. Lett.,* **41**(3), 850–857.
Omori, F., 1894. On aftershocks of earthquakes, *J. College Sci., Imperial University of Tokyo,* **7,** 111–200.
Reasenberg, P.A. & Jones, L.M., 1989. Earthquake hazard after a mainshock in California, *Science,* **243**(4895), 1173–1176.
Rhoades, D.A. & Evison, F.F., 2004. Long-range earthquake forecasting with every earthquake a precursor according to scale, *Pure appl. Geophys.,* **161**(1), 47–72.
Roberts, R.D., Anderson, A. R.S., Nelson, E., Brown, B.G., Wilson, J.W., Pocernich, M. & Saxen, T., 2012. Impacts of forecaster involvement on convective storm initiation and evolution nowcasting, *Weather Forecast.,* **27**(5), 1061–1089.
Roebber, P.J., 2009. Visualizing multiple measures of forecast quality, *Weather Forecast.,* **24**(2), 601–608.
Schorlemmer, D. *et al.*, 2018. The collaboratory for the study of earthquake predictability: achievements and priorities, *Seismol. Res. Lett.,* **89** , doi:10.1785/0220180053.
Scognamiglio, L. *et al.*, 2012. The 2012 Pianura Padana Emiliana seimic sequence: locations, moment tensors and magnitudes, *Ann. Geophys.,* **55,** 549–559.
Shebalin, P.N., Narteau, C., Zechar, J.D. & Holschneider, M., 2014. Combining earthquake forecasts using differential probability gains, *Earth Planets Space,* **66**(1), 1–14.
Stallone, A. & Falcone, G., 2021. Missing earthquake data reconstruction in the space-time-magnitude domain, *Earth Space Sci.,* **8**(8), e2020EA001481 , doi:10.1029/2020EA001481.
Stephenson, D., 2000. Use of the odds ratio for diagnosing forecast skill, *Weather Forecast.,* **15,** 221–232.
Taroni, M., Marzocchi, W., Schorlemmer, D., Werner, M.J., Wiemer, S., Zechar, J.D., Heiniger, L. & Euchner, F., 2018. Prospective CSEP evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for Italy, *Seismol. Res. Lett.,* **89**(4), 1251–1261.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram, *J. geophys. Res.: Atmos.,* **106**(D7), 7183–7192.
van der Elst, N.J., Hardebeck, J.L., Michael, A.J., McBride, S.K. & Vanacore, E., 2022. Prospective and retrospective evaluation of the U.S. Geological Survey Public Aftershock Forecast for the 2019â€2021 Southwest Puerto Rico earthquake and aftershocks, *Seismol. Res. Lett.,* **93**(2A), 620–640.
Vere-Jones, D., 1995. Forecasting earthquakes and earthquake risk, *Int. J. Forecast.,* **11**(4), 503–538.
Wiemer, S. & Katsumata, K., 1999. Spatial variability of seismicity parameters in aftershock zones, *J. geophys. Res.: Solid Earth,* **104**(B6), 13135–13151.
Wiemer, S., Gerstenberger, M. & Hauksson, E., 2002. Properties of the aftershock sequence of the 1999 Mw 7.1 Hector Mine Earthquake: implications for aftershock hazard, *Bull. seism. Soc. Am.,* **92**(4), 1227–1240.
Woessner, J., Christophersen, A., Zechar, J.D. & Monelli, D., 2010. Building self-consistent, short-term earthquake probability (step) models: improved strategies and calibration procedures, *Ann. Geophys.,* **53**(3), 141–154.
Zechar, J.D., 2010. Evaluating earthquake predictions and earthquake forecasts: a guide for students and new researchers, *Community Online Resource for Statistical Seismicity Analysis* , doi:10.5078/corssa-77337879.
Zechar, J.D. & Jordan, T.H., 2008. Testing alarm-based earthquake predictions, *Geophys. J. Int.,* **172**(2), 715–724.
Zechar, J.D. & Jordan, T.H., 2010. The area skill score statistic for evaluating earthquake predictability experiments, in *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II,* pp. 39–52, Springer.
Zechar, J.D., Gerstenberger, M.C. & Rhoades, D.A., 2010. Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. seism. Soc. Am.,* **100**(3), 1184–1195.
Zhuang, J., Ogata, Y. & Wang, T., 2017. Data completeness of the Kumamoto earthquake sequence in the JMA catalog and its influence on the estimation of the ETAS parameters, *Earth Planets Space,* **69,** 1–12.
Zhuang, J., Wang, T. & Kiyosugi, K., 2020. Detection and replenishment of missing data in marked point processes, *Stat. Sinica,* **30**(4), 2105–2130.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Table S1. Statistical tests and procedures performed in this paper to assess the reliability of the OEF-Italy forecasts in the first 7 yr of real-time operativity.

Table S2. Forecasts of the OEF-Italy system in which the number of observed target events is outside the 99 per cent CI obtained from 10 000 simulations. The symbol $\mathbb{P}$ stays for probability.

Table S3. Contingency table for the four classes of variables identifiable through the forecasting experiment. The positive observation '1' represents the event 'at least one target earthquake occurred in the spatio-temporal window considered'.

Table S4. ETAS (Lombardi & Marzocchi 2010) and ETES (Falcone *et al.* 2010) specific rate functions.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.